CMST SI(1) 119-124 (2010) DOI:10.12921/cmst.2010.SI.01.119-124

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **Special Issue 2010,** 119-124

PIONIER Network Digital Libraries Federation – Interoperability of Advanced Network Services Implemented on a Country Scale

Agnieszka Lewandowska, Marcin Werla

Poznań Supercomputing and Networking Center
Noskowskiego 12/14, 61-704 Poznań, Poland
e-mail: {jagna/mwerla}@man.poznan.pl

(Received: 25 August 2010; revised: 9 November 2010; published online: 23 November 2010)

Abstract: In March 2010 the Digital Library of the Wielkopolska Region, the largest and one of the oldest Polish digital libraries reached the level of 100 000 objects available on-line. Since its inauguration in 2002, the dynamics of the development of Polish digital libraries has increased significantly. The common effort of several hundreds of scientific and cultural institutions in Poland led to the present situation in which there are over 50 publicly available digital libraries and together they give access to around 425 000 digital objects. Such a big number of similar network services lead to the need of integration mechanisms allowing to achieve a synergy effect perceptible for end users as well as for digital library creators and international digital library initiatives. Such mechanism developed by PSNC and named PIONIER Network Digital Libraries Federation, has been available as a network service since June 22-nd, 2007 and can be reached at http://fbc.pionier.net.pl/. This paper shortly presents the history of development of Polish digital libraries, showing the factors that have influenced the initial form and advancement of the Federation. It also describes the Federation architecture, organizational aspects of its operation, and key challenges including the integration of Polish and European digital libraries infrastructures.

Key words: digital libraries, services interoperability, services infrastructure, metadata exchange, OAI-PMH protocol

I. INTRODUCTION

In October 2002, the Digital Library of the Wielkopolska Region (http://www.wbc.poznan.pl/) was made publicly available. It was the first Polish regional digital library and at the same time the first digital library based on the dLibra software (http://dlibra.psnc.pl/) developed by Poznań Supercomputing and Networking Center. Since its beginning, the library has been co-created by Poznań Foundation of Scientific Libraries and several other cultural and scientific institutions from Wielkopolska [1]. Nowadays this digital library is the largest one in Poland and gives access to over 100 000 digital objects. Two years after the Wielkopolska Digital Library premiere, another digital library in Poland was set up – the Digital Library of Wrocław University of Technology, later transformed into the regional Lower Silesia Digital Library. In the next five years, the dynamics of the creation of new digital libraries in Poland increased significantly and the joint effort of several hundreds of various institutions led to the present state of the PIONIER Network where there areover 50 digital libraries which give access to around 425 000 of digital objects.

This is a significant number of network services which are quite similar in many aspects, both functional and technical. What was missing was a mechanism allowing to integrate all these independently functioning services. When such mechanism was designed for the PIONIER Network, its precedent aim was defined as a synergy achieved between digital libraries distributed in Poland, allowing new possibilities to all involved parties including end-users, digital libraries creators and external systems and initiatives potentially interested in cooperation. The work on this service was initiated in PSNC in 2006 and the mission of the Digital Libraries Federation was defined as follows:

- facilitate the use of resources from Polish digital libraries,

- increase the visibility and popularity of resources from Polish digital libraries in the Internet,
- enable new advanced network services based on the resources from Polish digital libraries to Internet users and digital libraries creators.

It was decided that the main tool for implementing this mission would be be the aggregation and processing of metadata of objects available in Polish digital libraries and exposition of new services using this metadata [2]. The first version of the Federation had its premiere on 22-nd June, 2007 at the meeting of the PIONIER Consortium Supervisory Board in Kraków, Poland. Since this day the new network service has been made publicly available at http://fbc.pionier.net.pl/ [3].

The next section of this paper gives a general overview of the Federation and its current functionality. Later selected aspects of the Federation implementation are discussed, with a focus on technical details of the Federation metadata workflow. The fourth section depicts the experiences related to the cooperation between the Federation and Europeana (formerly: European Digital Library) and other similar initiatives. The paper ends with a summary and short discussion about the directions of future works.

II. PIONIER NETWORK DIGITAL LIBRARIES FEDERATION OVERVIEW

As mentioned above, the main mechanism of the Federation is the aggregation of the metadata from Polish digital libraries. This mechanism is based on the OAI-PMH protocol [4] and the Dublin Core Metadata Element Set (DCMES) [5] – a basic metadata schema developed for the description of electronic resources and required by the OAI-PMH protocol. The use of this protocol allows for automatic querying of all digital libraries compatible with it. This allows obtaining the information about somehow modified (i.e. added, changed or deleted) metadata records, where each record is metadata expressed in specific metadata schema and describing particular digital object. In case of the DCMES the record can consist of 15 various fields like title, creator, type, language, format, etc. As the communication between two interacting systems (called data provider and service provider) is fully automated, the period of data updates is limited in most cases only by the hardware performance. In case of the Federation the update is performed each night.

The services available for the Federation users are the following:

- Search (simple and advanced) – this functionality allows end users to search in all metadata aggregated in the Federation. The simple search allows users to select in the entire metadata record or in one specific metadata schema element (e.g., the title). The advanced search allows to combine several criteria connected with Boolean operators and optionally limited with an additional modification of date based restriction. The results of such simple or advanced query can be received as an HTML site, OAI-PMH set or RSS feed.
- Resolution of persistent identifiers – this functionality allows end users to use permanent links to the digital objects like:
  http://fbc.pionier.net.pl/id/oai:www.wbc.poznan.pl:8711
  in order to reference these objects. Because of the use of the OAI Id concept, the unique identifiers are generated automatically by the distributed digital libraries and the Federation just stores them and provides a permanent URL to access objects connected with these identifiers. An example advantage of using such references instead of direct ones like:
  http://www.wbc.poznan.pl/dobra/doccontent?id=8711
  is when the digital library management system of a particular digital library is upgraded or replaced, the direct link to the object can be changed. In such case the information about such link will be automatically updated in the Federation database and the persistent link realized via the federation will be still valid.
- Digitization plans, coordination of digitization – besides gathering the metadata of objects available on-line, the Federation also collects the information about digitization plans if these are available (currently around 10 000 records). This information is available for searching and browsing for the end users, but it is also used as a basis for the web service allowing to remotely detect if the metadata record submitted to this service is similar to any other object registered in the Federation – already digitized or planned for digitization. The similarity is defined by a custom made algorithm utilizing the Levenshtein Distance algorithm. This service is intensively utilized by the majority of Polish digital libraries and therefore is a semi-automatic mechanism supporting the coordination of digitization in Poland.

Besides the metadata-based services described above, the Federation also contains also a detailed database of all Polish digital libraries, including general descriptions of each library, geolocation data, list of cooperating institutions, news feeds, recommended object, etc. The Federation is also a gateway to the networked user profile feature available in

several Polish digital libraries. This feature allows users to use single profile in several digital libraries and it is implemented as an extension to the Shibboleth AA model.

At the moment the most challenging aspects of the implementation of such country-scale automated services cooperation are connected with the process of metadata aggregation and processing. The selected details of this process are described in the next section.

III. SELECTED ASPECTS OF THE FEDERATION IMPLEMENTATION

The main interoperability problem concerning the implementation of the Digital Libraries Federation is connected with the metadata interoperability. Theoretically the use of the OAI-PMH protocol forces the use of a common metadata schema – the Dublin Core Metadata Element Set. But this is just a basic technical interoperability. To expose the metadata in the DCMES, the majority of digital libraries in Poland have to transform each record from their internal metadata schema. Such transformation/mapping is executrd on the basis of automated rules and can lead to the loss of information if the internal schema is much richer than the DCMES (like MARC standard for example). Such transformation is the first step of the metadata workflow in the network of Polish digital libraries.

The second step is the transfer of the metadata from particular digital library to the Federation. As it was mentioned, this is performed each night, by the means of the OAI-PMH protocol. The harvested metadata is validated (for example, if it contains URL to the described digital object), stored in the Federation database and indexed for searching purposes.

The next step is the automated enrichment of the stored metadata in several ways. The Federation software tries to download a thumbnail of each digital object if possible, or if the thumbnail is not delivered by the source digital library, it tries to prepare such thumbnail on the basis of the original object. Besides, the metadata record is also analysed in order to detect if the object is publicly available (99% of objects) or if the access is somehow restricted. This metadata analysis also aims to assign the object into one of four categories: TEXT, IMAGE, AUDIO and VIDEO. Such enriched metadata is ready to be exposed to the Federation end-users.

Additionally to the needs of external services which use the metadata collected by the Federation, the metadata may be optionally normalized. For example, the DCMES language element can be transformed to a form compatible with the ISO 639-2 standard [6]. The normalization is dependent on the service which harvests the metadata from the Federation and it can also be connected with an advanced process of the selection of records which have to be exposed for a particular service. This is described in more detail in the next section.

IV. COOPERATION WITH THE EUROPEANA AND OTHER INITIATIVES

As the number of digital objects in Polish digital libraries was increasing and their diversity became more impressing, the important step in the advancement of the Federation services was joining those materials with other pan-European initiatives. Hence the Federation is fulfilling its mission and promoting Polish publications in Europe. In the next few paragraphs the connection with the European digital libraries infrastructure is described in more detail.

The best known initiative the Federation is cooperating with is the Europeana – Europeana Digital Library, Museum and Archive (http://europeana.eu). It is a flagship project of European Union aiming at giving everyone access to European cultural heritage which is mostly in the analog form today. As Europeans live in a digital age, access via the Internet to the cultural heritage is a must and it is expected to boost European economy, education, science, art, etc. In the mid-2010, Europeana is going to present about 10 million diverse digital objects from European countries: not only books, journals, but also recordings, films, painting, sculptures and many more.

Europeana is built by a group of EU-funded projects, each of them focusing on a different aspect of realizing such an ambitious project. The project focusing on bringing regional and local content to European is EuropeanaLocal. Poznań Supercomputing and Networking Center is its regional content coordinator for Poland. EuropeanaLocal, as well as Europeana, is strongly supporting aggregators and its infrastructure in Europe. Aggregators are considered crucial partners because they facilitate contact with content providers, assure data conformance and quality, promote Europeana as an initiative and hence bring new content providers to Europeana. In short, they overload work Europeana has to do. As the resources of Europeana are limited, only strong cooperation with aggregators assures its sustainability and extension. To express its support Europeana has published an “Aggregators Handbook” [7] as a way of promoting aggregators creation. Fortunately it is worth noting that the document presents the PIONIER Network Digital Libraries Federation as a successful case study of a national aggregator.

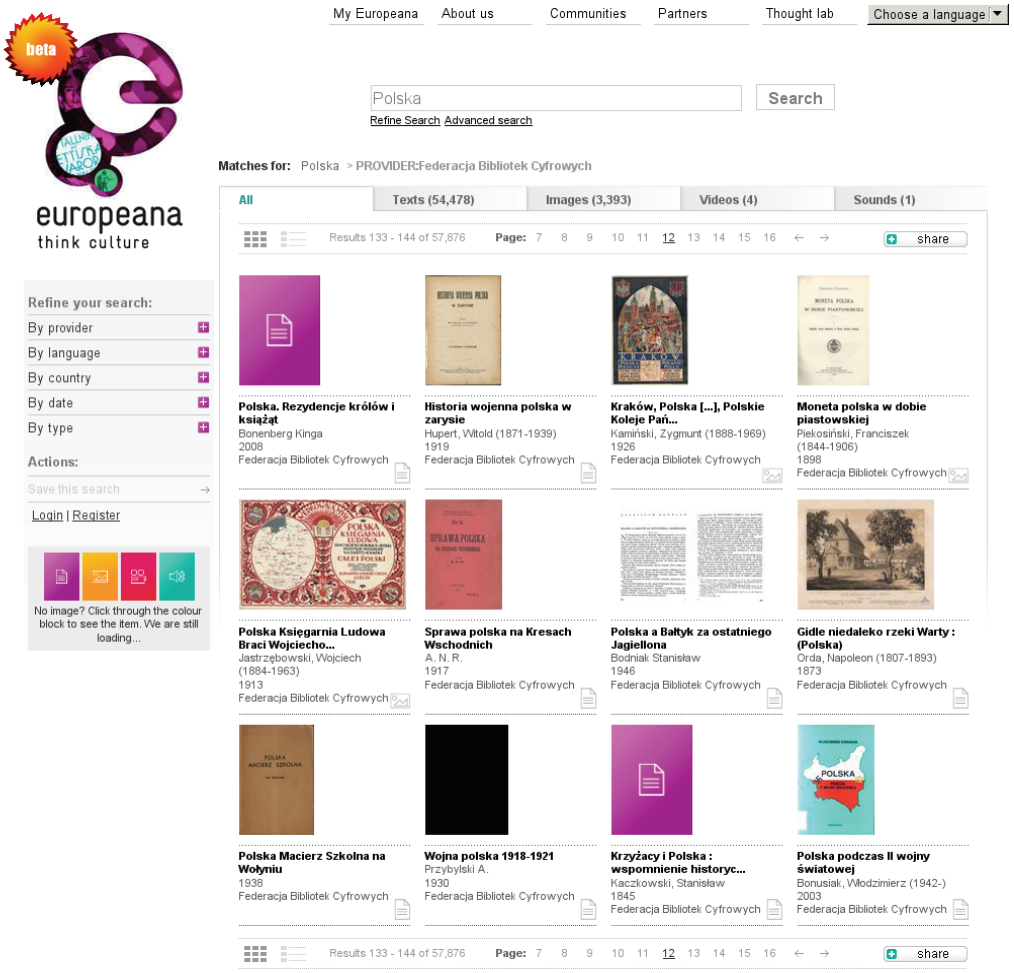

Fig. 1. Content transferred via PIONIER DLF in Europeana portal

The first transfer of content from the Federation to Europeana happened on 11th of December, 2009. Over 250 000 Polish digital objects from more than 40 local and regional digital libraries were ingested. It resulted in increasing the percentage of the Polish content in Europeana from less than 1% to 5.5%! In order to transfer its data, the Federation had to meet the Europeana requirements: all objects should be publicly available and must be described in Europeana Semantic Elements metadata scheme [8]. Moreover, some guidelines over mapping and standardization were recommended [9]. On the other hand we added our own restriction; we wanted to transfer data of only those digital libraries which agreed to it.

The first step was to create a mechanism for transferring metadata further. The improved OAI-PMH interface was implemented with a possibility to configure the metadata of which content providers is transferred further. As this phase ended, the metadata adjustments was started. The first obstacle was the difference between metadata schemes. The Federation gathers data in DCMES, but Europeana uses a different description standard, Europeana Semantic Elements. A semi-automatic mechanism of mapping between attributes of DCMES and ESE was created. In order to do that, some static files with mappings definitions, which are later used for the automatic transformation between description standards were prepared. Furthermore, according to the Europeana guidelines, one of the messiest attributes, a language field was cleaned and normalized. In Europeana portal it should be presented in the ISO 639-2 standard. After that the last requirement from Europeana was selecting publicly available objects. As the Federation keeps only objects metadata, the selection process was based on it. Description values indicating restricted access were picked and used to automatically select public objects. Next our content providers were contacted and asked for permission to transfer their data further. After receiving substantial number of positive responses, our data were checked in the

Content Checker, a Europeana test portal for examining metadata correctness and its presentation. During this test phase some minor corrections were added as well. The final part was contacting with Europeana team and asking for ingestion. The results are presented in Fig. 1. Nowadays the Federation transfers almost 350 000 digital objects to the Europeana from more than 50 digital libraries.

DART-Europe (http://www.dart-europe.eu) is the second initiative to which the Federation forwards the gathered metadata. It aims at giving access to electronic theses and dissertations (ETDs) from Europe and enables researchers to discover European ETDs. It requires that all collected digital object are open-access, full-text and on the research level. Nowadays it presents researchers' papers from 11 European countries and almost 300 universities. Poznań Supercomputing and Networking Center is an official partner of DART-Europe. What distinguishes this aggregator from Europeana is a specific set of requirements. They are not looking for digital content on any type, but are focusing efforts only on a small, well-defined subset of such objects. Thanks to it, they are able to distinct important information from collected metadata, which is specific for a declared type of objects. For example, due to the importance of up-to-date data in science, the DART-Europe requires that metadata contains a date of ETDs creation. Furthermore, they cooperate with each data provider to distinct granting institution from the ETDs description. Therefore, it is attributing the research institutions for its work.

The first connection of the Federation with DART-Europe occurred on 20-th January 2010. The Federation pushed descriptions of 256 ETDs, from 13 universities and 6 digital libraries. The preparation process was focusing on ETDs identification in Polish digital libraries, which was realized with in strong cooperation with content providers. Moreover, the proper transfer mechanism was created – the OAI-PMH interface was expanded with searching capabilities while requesting OAI-PMH sets (so called dynamic sets [10]). Because the initial connection was well received among content providers, it resulted in a substantial increase of transferred objects. Nowadays DART-Europe presents 1 227 ETDs gathered via the Federation. They come from 31 universities and 17 digital libraries from Poland.

V. SUMMARY

In this paper we have described the Digital Libraries Federation built on the top of digital libraries deployed in the PIONIER Network. This Federation is a key integration point for tenths of digital libraries in Poland, a single point of discovery and access to the resources of these libraries, and also a very convenient source of data for everyone interested in the metadata of objects published on-line by Polish cultural and scientific institutions.

The European initiatives like the Europeana and DART-Europe mentioned in the previous section, technical requirements and recommendations developed by these initiatives create new challenges for the Federation development, connected not only with the technical interoperability but also with the semantic and cross-language issues. Another source of new tasks is the lack of well-defined standards adapted for use in Poland, which could be used for the creation and integration of metadata for digital representations of cultural heritage objects coming from multiple domains (libraries, museums, archives).

The development plans for the Federation are twofold. First of all, it is considered to prepare a software package based on the Federation toolkit allowing easy deployment of such services in other countries or regions. But the transformation of an in-house service to a user-friendly software product requires a lot of effort. The other direction of development of the Federation is focused on the automated metadata processing and enrichment. As the open linked data movement (http://linkeddata.org/) is gaining its momentum, the next steps of the Federation evolution will for sure reflect the most important aspects of this idea.

References

[1] M. Górny, P. Gruszczyński, C. Mazurek, J.A. Nikisch, M Stroiński, A. Swędrzyński, *Zastosowanie oprogramowania dLibra do budowy Wielkopolskiej Biblioteki Cyfrowej*, in*: Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej. Seria „Technologie Informacyjne",* Wydawnictwo Politechniki Gdańskiej 109-117 (2003).

[2] C. Mazurek, M. Stroiński, J. Węglarz, M. Werla, *Metadata harvesting in regional digital libraries in PIONIER Network*, Campus-Wide Information Systems 23 (4), 241-253 (2006).

[3] M. Kosiedowski, C. Mazurek, M. Stroiński, M. Werla, M. Wolski, *Federating Digital Library Services for Advanced Applications in Science and Education.* Computational Methods in Science and Technology 13, 101-112 (2007).

[4] C. Lagoze, H. Van de Sompel, *The Open Archives Initiative Protocol for Metadata Harvesting*. accessed from http://www.openarchives.org/OAI/openarchivesprotocol.html Open Archives Initiative (2004).

[5] *The Dublin Core metadata element set,* http://dublincore.org/documents/dces/.

[6] *The ISO 639-2 standard,* http://www.loc.gov/standards/iso639-2/.

[7] *Aggregators Handbook*, accessed from: http://version1.europeana.eu/c/document_library/get_file?-uuid=94bcddbf-3625-4e6d-8135-c7375d6bbc62&groupId-=10602, The Europeana Team (2010).

[8] *Europeana Semantic Elements specifications v3.2.2*, accessed from: http://version1.europeana.eu/c/document_library/get_file?uuid=c56f82a4-8191-42fa-9379-4d5ff8c4ff75&groupId=10602, The Europeana Team (2010).
[9] *Metadata Mapping and Normalisation Guidelines*, accessed from: http://version1.europeana.eu/c/document_library/get_file?uuid=58e2b828-b5f3-4fe0-aa46-3dcbc0a2a1f0&groupId=10602, The Europeana Team (2010).
[10] C. Mazurek, M. Mielnicki, M. Werla, *Extending OAI-PMH protocol with dynamic sets definitions using CQL language. White paper*, accessed from: http://dl.psnc.pl/biblioteka/dlibra/docmetadata?id=254 (2009).

AGNIESZKA LEWANDOWSKA graduated in Computer Science at Poznań University of Technology. Since 2007 she has been working as a computer system analyst in PSNC Digital Libraries Team. She is the main programmer and administrator of the Digital Libraries Federation. Her professional interests include software engineering and selected apects of data analysis.

MARCIN WERLA, M.Sc. Eng., is the leader of the Digital Libraries Team in the PSNC Network Services Department. They focus on the development of dLibra Digital Library Framework, Digital Libraries Federation service and also participate in several European projects. Marcin is also the main organizer of the annual national "Digital Libraries" workshop and "Polish Digital Libraries" Conference held by PSNC. He has authored or co-authored of several papers in professional journals and conference proceedings.